МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО”

Савастьянов Володимир Володимирович

УДК 004.82:005.52:519-7.51

СУПРОВОДЖЕННЯ ПРОЦЕСУ ПЕРЕДБАЧЕННЯ З НАЯВНІСТЮ СЛАБКО СТРУКТУРОВАНИХ ДАНИХ ЗАСОБАМИ ТЕКСТОВОЇ АНАЛІТИКИ

Спеціальність 01.05.04 – системний аналіз і теорія оптимальних рішень

АВТОРЕФЕРАТ

дисертації на здобуття наукового ступеня

кандидата технічних наук

Київ – 2021

Дисертація є рукопис

Робота виконана у Інституті прикладного системного аналізу (ІПСА) Національного технічного університету України “Київський політехнічний інститут імені Ігоря Сікорського”.

Науковий керівник: Чл.-кор. НАН України, доктор технічних наук, професор
Панкратова Наталія Дмитрівна,
Національний технічний університет України “Київський політехнічний інститут імені Ігоря Сікорського”,
заступник директора з наукової роботи ІПСА.

Офіційні опоненти: доктор технічних наук, професор,
Ланде Дмитро Володимирович,
Інституту проблем реєстрації інформації НАН України,
завідувач відділу спеціалізованих засобів моделювання;

кандидат технічних наук, старший науковий співробітник
Колос Людмила Миколаївна,
Інститут космічних досліджень НАН України та ДКА України, старший науковий співробітник.

Захист відбудеться 14 травня 2021 о 15 годині 00 хвилин на засіданні спеціалізованої вченої ради Д 26.002.03 при Національному технічному університеті України “Київський політехнічний інститут імені Ігоря Сікорського” за адресою: 03056, м.Київ - 56, пр. Перемоги, 37, корп. №35, ауд. 001.

З дисертацією можна ознайомитися у бібліотеці Національного технічного університету України “Київський політехнічний інститут імені Ігоря Сікорського” за адресою: 03056, м. Київ - 56, пр. Перемоги, 37.

Автореферат розіслано 14 квітня 2021 р.

Вчений секретар

спеціалізованої вченої ради Д 26.002.03

д.ф.-м..н., професор В.О.Капустян

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Зумовлюється стрімким розвитком технологій, надходженням та накопиченням інформації (у вигляді слабко структурованих даних та знань), їх впливом на оточуюче середовище, що пов'язано з необхідністю розробки апарату математичного забезпечення супроводження процесу передбачення з використанням прийомів та методів текстової аналітики.

Дисертаційна робота присвячена розробці та застосуванню прикладної наукової методології системного аналізу для супроводження задач передбачення (у роботах М.З. Згуровского, Н.Д. Панкратової). До цього часу передбачення існувало у вигляді наборів методів, що були об’єднані організаційними процедурами, або системами підтримки організаційних процедур (рекомендації UNIDO, Handbook of Knowledge Society Foresight).

У сучасних state-of-art роботах текстова аналітика застосовується у передбаченні все більш і більш часто (роботи Ozcan Saritas, Serhat Burmaoglu, Kayser V., Blind K., Dreher C). Текстова аналітика дозволяє обробляти великі обсяги слабко структурованих даних, вилучати об’єкти чи формувати структуру досліджуваного об’єкту.

Проте, текстова аналітика конкурує з іншими методами якісного аналізу, або формує опис предметної галузі у деякому оптимальному вигляді (онтології предметної галузі) (в роботах Т.А. Гаврилової, В.А. Горової, Е.С. Болотникової, А.В. Палагина, Н.Г. Петренко). Це зумовлено тим, що зменшення невизначеності за рахунок оброблення більших обсягів вхідних даних для формування онтології предметної галузі, чи вилучення асоціативних зв’язків об’єктів, суб’єктів або систем для формування переліків ключових технологій, є дуже важливий етап розвитку технологій передбачення у сучасному темпі росту обсягів знань.

Тобто, зменшення невизначеності є однією з важливих проблем процесу передбачення (ПП). А тому застосування текстової аналітики на всьому процесі передбачення замість її використання на окремих інтервалах часу окремим методом, є суттєвим удосконаленням при оцінюванні досяжності мети процесу передбачення. У представленій роботі текстова аналітика використовується у принципово новій ролі – для супроводження ПП (СПП) через систематичне зменшення невизначеності через неперервну структуризацію предметної області, вилучення введених у процес супроводження метаданих та вимірювання показників інформованості відносно структури знань, вхідних документів та метаданих, вилучення об’єктів, цілей, проблем, трендів та ін. з метою забезпечення методів якісного аналізу достовірними вхідними даними, що обумовлює актуальність дисертаційного дослідження.

Зв’язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконана у відділі Математичних методів системного аналізу Інституту прикладного системного аналізу Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського» у відповідності до планів науково-дослідних робіт: «Розробка та дослідження теоретичних основ методології сценарного аналізу», № держреєстрації 0107U004124, 2007-2011 рр.; «Розробка платформи сценарного аналізу в межах сталого розвитку», № держреєстрації 0110U002364, 2010–2011 рр.; «Розробка інформаційної системи

супроводження процесу передбачення», № держреєстрації 0112U003164, 2012 -2013 рр.; «Розробка теоретичних засад прийняття рішень на основі методології передбачення», № держреєстрації 0112U000558, 2012-2016 рр.; «Розробка інформаційно-аналітичних засобів дослідницької служби у складі інтегрованої інформаційно-аналітичної системи "Електронний Парламент"», 2012-2013 рр.; «Синтез методологій передбачення і когнітивного моделювання щодо розробки стратегії інноваційного розвитку регіону», № держреєстрації 0114U004076, 2014-2015 рр.; «Розробка інформаційно-експертної системи передбачення з урахуванням поглибленої аналітики неструктурованих даних», № держреєстрації 0114U001533, 2014-2015 рр.; «Modeling and Mitigation of Social Disasters Caused by Catastrophes and Terrorism» NUKR.SFPP G4877, 2015-2018 рр.; «Побудова інформаційно-аналітичної платформи сценарного аналізу на основі великих обсягів слабко структурованої інформації»: № держреєстрації, 0118U003779, 2018–2020 рр.; «Розроблення теоретичних засад сценарного аналізу на основі великих обсягів слабко структурованої інформації», № держреєстрації 0115U002499, 2017-2021 рр.

Метою дослідження є розробка математичного забезпечення супроводження процесу передбачення з наявністю слабко структурованих даних засобами текстової аналітики.

Для досягнення мети дослідження поставлені й розв'язані такі **задачі**:

- проаналізувати існуючі підходи до СПП;
- розробити концепцію СПП;
- розробити системний підхід до СПП на основі прийомів обробки слабко структурованих даних і текстової аналітики:
    - розробити інформаційну модель ПП із наявністю слабко структурованих даних та його метадані;
    - запропонувати інформаційну модель предметної галузі;
    - створити концептуальну модель якості знань;
    - розробити модель та підходи щодо вилучення фактів та знань із слабко структурованих даних;
    - розробити модель для врахування емоційної забарвленості;
    - розробити прийоми та алгоритми щодо вилучення та аналізу об'єктів-метаданих інформаційної моделі передбачення та їх властивостей;
    - дослідити ситуацію виникнення конфліктів знань та створити прийоми щодо їх розв'язання;
- провести апробацію інформаційної моделі СПП із наявністю слабко структурованих даних;
- створити обчислювальні модулі реалізації прийомів обробки слабко структурованих даних і текстової аналітики;
- застосувати зазначений системний підхід до реалізації кейсів передбачення.

Об'єктом дослідження є побудова супроводження процесу передбачення щодо складних систем різної природи.

Предмет дослідження - моделі, методи, прийоми, методологія системного аналізу, методологія процесу передбачення, засоби текстової аналітики.

Методи дослідження:

- Методологія сценарного аналізу - для вивчення процесу застосування окремих методів у певній послідовності із встановленням визначених взаємозв'язків між ними.
- Теорія прийняття рішення - для дослідження закономірностей вибору раціональних альтернатив в умовах конфліктуючих цілей та багатофакторних ризиків.
- Методи якісного аналізу - як складові системної методології передбачення для обґрунтування експертних суджень.
- Текстова аналітика - для отримання інформації, фактів, емоціонального забарвлення, суджень, зв'язків та знань з наборів слабко структурованих даних, у тому числі текстових документів природною мовою, при застосуванні методів інтелектуального аналізу даних.

Наукова новизна отриманих результатів. Виконані у дисертаційній роботі дослідження дозволили отримати такі теоретичні та практичні результати:

Уперше:

- запропоновано концепцію супроводження процесу передбачення;
- розроблено системний підхід до супроводження процесу передбачення, що відрізняється від існуючих застосуванням прийомів обробки слабко структурованих даних на основі текстової аналітики;
- розроблено інформаційну модель супроводження процесу передбачення, що відрізняється від існуючих урахуванням наявності великих обсягів слабко структурованих даних;
- введено показники інформованості щодо виконання процесу передбачення, що було вперше введено у дисертаційній роботі;
- розроблено формалізацію прийомів для вилучення знань (метаданих) з наявністю слабко структурованих даних: вилученням об'єктів та їх властивостей з використанням існуючого словаря позитивних або негативних слів; вилученням позитивних або негативних слів з використанням існуючої таксономії об'єктів та їх властивостей; визначенням значимості іменних груп щодо бажаних і небажаних фактів; ідентифікацією фактів потенційно позитивного або негативного показника; розкриттям неоднозначності ситуацій зміни емоційно-семантичної орієнтації.

Удосконалено:

- модель обробки слабко структурованих даних, що відрізняється можливістю вилучення фактів з текстів, у тому числі з урахуванням емоційно-семантичної орієнтації;
- прийоми створення метрики для врахування емоційної забарвленості, що відрізняються запропонованим коефіцієнтом зважування емоцій на корпусі документів із урахуванням інтервалу часу.

На базі теоретичної частини створено автоматизовані модулі реалізації прийомів обробки слабко структурованих даних і текстової аналітики.

Зазначений підхід викладено у вигляді ряду кейсів щодо супроводження процесу передбачення.

Практичне значення одержаних результатів полягає у створенні формалізованої, теоретично обґрунтованої стратегії супроводження процесу

передбачення з метою зменшення невизначеності при розв’язанні практичних задач створення бажаного майбутнього.

Розроблено алгоритми обробки вхідних даних та процес оцінювання якості знань та визначення конфліктів знань на основі інформаційної моделі супроводження процесу передбачення. Для обробки вхідних даних на базі моделі вилучення фактів із текстів природною мовою розроблено лексичні обмеження у вигляді шаблонів правил, що дозволяють будувати класифікатори предметної галузі. Створено комплексні правила для вилучення кожного типу знань/метаданих передбачення з слабко структурованих джерел.

Класифіковано типи вхідних даних. Створено алгоритм процесу обробки вхідної інформації при надходженні нових знань; описано функціонування додаткових блоків інформаційної моделі процесу передбачення. Створено ряд класифікаторів за галузями та згенеровано правила для класифікації автоматизованими засобами. Сформовані моделями правила та категоризатори є міжгалузевими та універсальними, а тому можуть застосовуватися у подальших дослідженнях з мінімальною модифікацією.

Створено програмні продукти на мові Python, адаптовані для використання як в рамках проектів з відкритим ПЗ (OpenSource), так і пропрієтарним (SAS(R)). Запропоновано схему масштабування програмного рішення.

На основі запропонованого системного підходу виконано низку практичних задач по замовленню Міністерств та відомств: Розробка платформи сценарного аналізу в межах сталого розвитку; Розробка інформаційної системи супроводження процесу передбачення для побудови логістики ПАТ "АрселорМіттал Кривий Ріг"; Розробка інформаційно-аналітичних засобів дослідницької служби у складі інформаційно-аналітичної системи “Електронний Парламент”; Modeling and Mitigation of Social Disasters Caused by Catastrophes and Terrorism та інші.

Запропонований у роботі системний підхід щодо супроводження процесу передбачення, разом із розробленими правилами та категоризаторами, накопиченими масивами знань, оглядом літературних джерел є корисними як методичний матеріал при написанні курсових та дипломних робіт, а також при складанні лекційних курсів з “Основи системного аналізу”, “Текстова аналітика” та ін.

Результати дисертаційної роботи впроваджені в навчальний процес кафедри математичних методів системного аналізу ІПСА КПІ ім. Ігоря Сікорського.

Особистий внесок здобувача. Всі наукові результати, що складають основний зміст роботи та становлять наукову новизну, отримані автором особисто. Зокрема, розроблено, концепцію супроводження процесу передбачення, теоретично обґрунтовано та практично апробовано модель з урахуванням показників інформованості та системний підхід до супроводження процесу передбачення з наявністю слабко структурованих даних засобами текстової аналітики.

У працях, написаних у співавторстві, здобувачеві належать: у праці [1] інформаційна модель, зручною для подання в пам'яті ЕОМ, що утворює базу і поле знань, побудована на основі мережі фреймів; стратегія інформаційного моделювання альтернатив сценаріїв, у праці [5] здобувачем розроблено нові метадані процесу передбачення, інформаційну модель процесу передбачення, алгоритм процесу обробки вхідної інформації, модифіковану модель вилучення фактів з текстів, у

праці [6] запропоновано метод обробки слабко структурованих даних у формуванні альтернатив сценаріїв, у праці [7] модифіковано інформаційну модель передбачення, що утворює базу знань, побудована на основі мережі фреймів; стратегія інформаційного моделювання альтернатив сценаріїв, у праці [8] синтез правил обробки вхідних даних у слабко формалізованому вигляді для вилучення факторів, концептів, причинно-наслідкових зв’язків, у праці [9] інформаційно-лексична модель соціально-економічної системи для категоризації даних, підходи текстової аналітики для вилучення трендів, фактів росту/падіння потенціально позитивного/негативного показника, у праці [11] алгоритм імпорту даних, що реалізує пошук пов'язаної інформації із зовнішніх джерел в режимі автоматичного пошуку за критеріями автоматичної агрегації з відомих джерел та напівавтоматичної агрегації з інших джерел інформації.

Апробація результатів дисертації. Наукові та практичні результати доповідались на семінарах та наукових конференціях:

- міжнародних наукових конференціях «Системний аналіз та інформаційні технології» SAIT (м. Київ, 2007 – 2017 рр);
- всеукраїнської науково-практичної конференції “Информационно-компьютерные технологии в экономике, образовании и социальной сфере” (м. Сімферополь, 2010).
- міжнародній науковій конференції “Интеллектуальные системы в информационном противоборстве”, (м. Москва, 2015)
- міжнародній науковій конференції IEEE UKRCON (м. Львів, 2019).

Публікації. Основні результати дисертаційної роботи опубліковано в 10 наукових працях, серед них 6 статей у наукових фахових виданнях (серед яких 2 статті у виданнях іноземних держав (Болгарія, Німеччина), 1 стаття у виданнях України, що включено до міжнародних наукометричних баз даних), 14 тез у матеріалах доповідей міжнародних і всеукраїнських конференцій.

Структура та обсяг дисертації. Дисертація складається зі вступу, переліку умовних позначень, чотирьох основних розділів, висновків, списку використаних джерел і додатків. Робота викладена на 203 сторінках і містить 122 сторінок основної частини, 48 рисунків, 12 таблиць і список використаних джерел із 130 найменувань.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтовано актуальність розробки супроводження процесу передбачення засобами текстової аналітики для слабко структурованих даних; визначені мета, об’єкт, предмет і методи дослідження; показано зв'язок з науковими програмами, планами; наведено наукову новизну та практичне значення одержаних результатів, висвітлено особистий внесок здобувача.

У **першому розділі** наведено огляд літератури за тематикою даної роботи та спорідненими питаннями; висвітлені результати, які були отримані іншими авторами. Зокрема, наведено огляд досліджень, що стосуються методології передбачення, методів якісного аналізу, інформаційної платформи сценарного аналізу. Проаналізовано також роботи, в яких пропонуються, аналізуються і

застосовуються методи текстової аналітики у передбаченні для вирішення задач предметної області. Додатково розглянуто роботи, в яких пропонується застосування моделей вилучення фактів щодо об’єктів та їх властивостей, моделей зважування емоцій, моделей для побудови онтологій предметних областей, у тому числі у передбаченні, що свідчить про велике значення цих досліджень для вирішення практичних задач та їх актуальність на сучасному етапі розвитку науки. Незважаючи на значні обсяги досліджень, присвячених вказаним тематикам, та значні досягнення у розвитку як теоретичних, так і практичних аспектів, існує ряд проблем, що потребують вирішення. Ці проблеми пов’язані із необхідністю застосування нових концепцій, ідей, моделей та алгоритмів, які здатні більш досконало враховувати особливості процесу передбачення, враховувати великі обсяги слабко структурованих вхідних даних, враховувати показники інформованості відносно росту знань, а, отже, зменшувати невизначеність знань у ході процесу передбачення. Отже, проведений огляд сучасної літератури на тему дисертації дозволяє аргументувати актуальність та практичну важливість проведених у роботі досліджень.

У **другому розділі** наведено системний підхід до супроводження процесу передбачення (ПП) засобами текстової аналітики для слабко структурованих даних, що складається з чотирьох етапів. Структурна схема системного підходу до супроводження процесу передбачення наведена на рис. 1.

На першому етапі визначається концептуальна модель супроводження ПП. Формується уява про процес та горизонт передбачення. Визначаються фактори росту та зменшення невизначеності горизонту передбачення.

Вводиться інформаційна модель ПП, у складі якої визначаються та вводяться додаткові базові інформаційні одиниці - метадані ПП. Визначається природа джерел інформації ПП у вигляді слабко структурованих даних природною мовою.

Запропоновано інформаційну модель предметної галузі - представлення предметних областей з використанням теоретико-множинних понять загальної теорії систем. Вводяться обмеження на зв’язки інформаційної моделі ПП, розглядаються варіанти представлення знань у вигляді ієрархічного класифікатору або онтології, окреслено переваги та недоліки. Розглянуто концепцію існування знань у часі. Введено інтегровані показники інформованості в залежності від часу для вимірювання змін у базі знань з часом та/або в залежності від обсягів надходження нових знань.

Пропонується використовувати представлення нових знань як класифікованих метаданих, при цьому самі класифікатори розробляються, доповнюються та використовуються повторно в рамках всіх інших сесій супроводження ПП.

Рис. 1. Структурна схема системного підходу до супроводження процесу передбачення засобами текстової аналітики для слабко структурованих даних

На всьому протязі супроводження ПП у рамках системного підходу постійно розраховуються та аналізуються показники інформованості.

На другому етапі вводиться та застосовується моделі та прийоми вилучення знань з текстів природною мовою. В рамках роботи модифіковано загальну модель вилучення фактів з текстів природною мовою для вилучення метаданих інформаційної моделі ПП та введено універсальні лексичні шаблони-обмеження для складання більш потужних правил вилучення метаданих. Моделі використовуються в рамках супроводження ПП для побудови прийомів та засобів обробки нових предметних областей та типів знань.

Створено прийоми щодо вилучення об'єктів предметної галузі для побудови та розширення класифікаторів, також прийоми для генерації класифікуючих правил для вузлів класифікаторів. Введено прийоми до обробки фактів, що містять потенційно позитивні та негативні показники, в тому числі з урахуванням плину часу та зміни контексту. Розглянуто ситуації конфліктів знань через зміну емоційно-семантичної орієнтації та прийоми до їх усунення.

На третьому етапі запропоновано інформаційну модель супроводження ПП. Визначаються класи вхідних даних, вводяться метадані для первинного анотування та метадані для супроводження ПП. Представлено алгоритм перетворення вхідних даних у метадані, показники інформованості та підходи усунення протиріч знань у базі знань. Розглянуто дані на виході супроводження ПП та можливості щодо їх застосування на різних етапах передбачення та у методах якісного аналізу.

На четвертому етапі проводиться адаптація та масштабування модулів обробки слабко структурованих даних у складі системи супроводження ПП з наявністю

слабко структурованих даних. На ряді кейсів показано застосування системного підходу щодо супроводження ПП з наявністю слабко структурованих даних засобами текстової аналітики.

Розроблений системний підхід застосовується на всьому життєвому циклі сесії передбачення. Створені на виході процесу супроводження артефакти (класифікатори, лексичні обмеження, правила, знання) можуть бути застосовані у наступних сесіях передбачення.

Концептуальна модель супроводження ПП розглядається у конусі часу (рис. 2). Сценарій у начальний момент часу T(0) має одну визначену ситуацію, що склалася, проте може мати декілька кінцевих ситуацій у майбутньому проміжку часу T(N). Якщо можливо передбачити основні ключові якісні та кількісні зміни, ПП залишається у просторі релевантних знань, та можливо з високою вірогідністю створити альтернативу сценарію з максимально правдоподібною ситуацією в майбутньому. Проте, чим менше ключових подій передбачено та відслідковується, тим більше вірогідність попасти у нерелевантну, спекуляційну область. Тобто, проблема відслідковування подій, трендів, ситуацій, визначення ключових технологій, накопичення та збереження важливих зв'язків, актуалізація релевантності з впливом часу - все це є невід'ємною частиною ПП при моделюванні сценаріїв майбутнього.

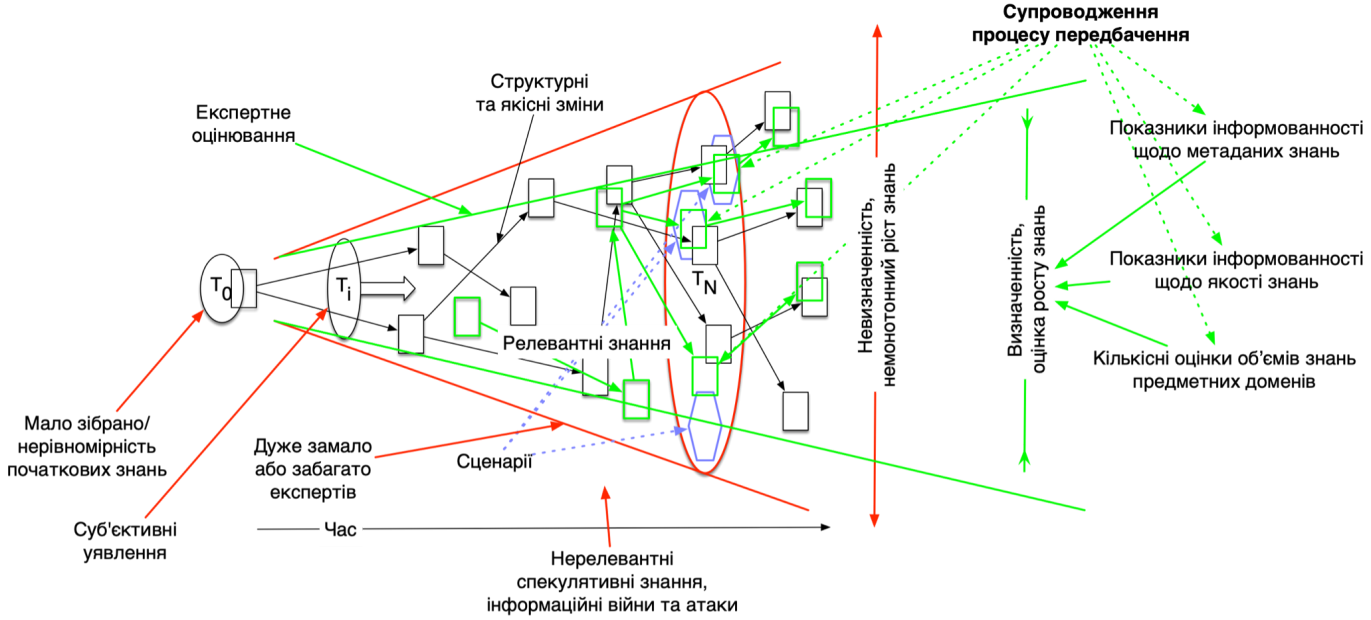

Рис. 2. Структура схема концептуальної моделі: конус часу та релевантність сценаріїв передбачення

Ще не набуті знання, екологічні нещастя, процеси, що застосовують ключові технології та економічні кризи, та вплив росту неякісної або викривленої інформації розширюють конус, збільшуючи невизначеність ситуації у момент часу T(N) у майбутньому. На звуження конусу впливають фактори, що зменшують невизначеність. Наприклад, через використання методів якісного аналізу та через супроводження процесу набуття та використання знань (введення показників інформованості) відповідно до їх структури, типів та джерел. Базові інформаційні одиниці передбачення розділено на метадані 1го та 2го рівнів: метадані описові (категорії) і метадані для логічних обчислень (факти та думки).

До інформаційної моделі ПП введено додаткові модулі: текстової аналітики, оцінки якості інформації, супроводження ПП. Також введено *додаткові метадані* до Бази знань. Структурна схема модифікованої інформаційної моделі ПП наведено на рис. 3.

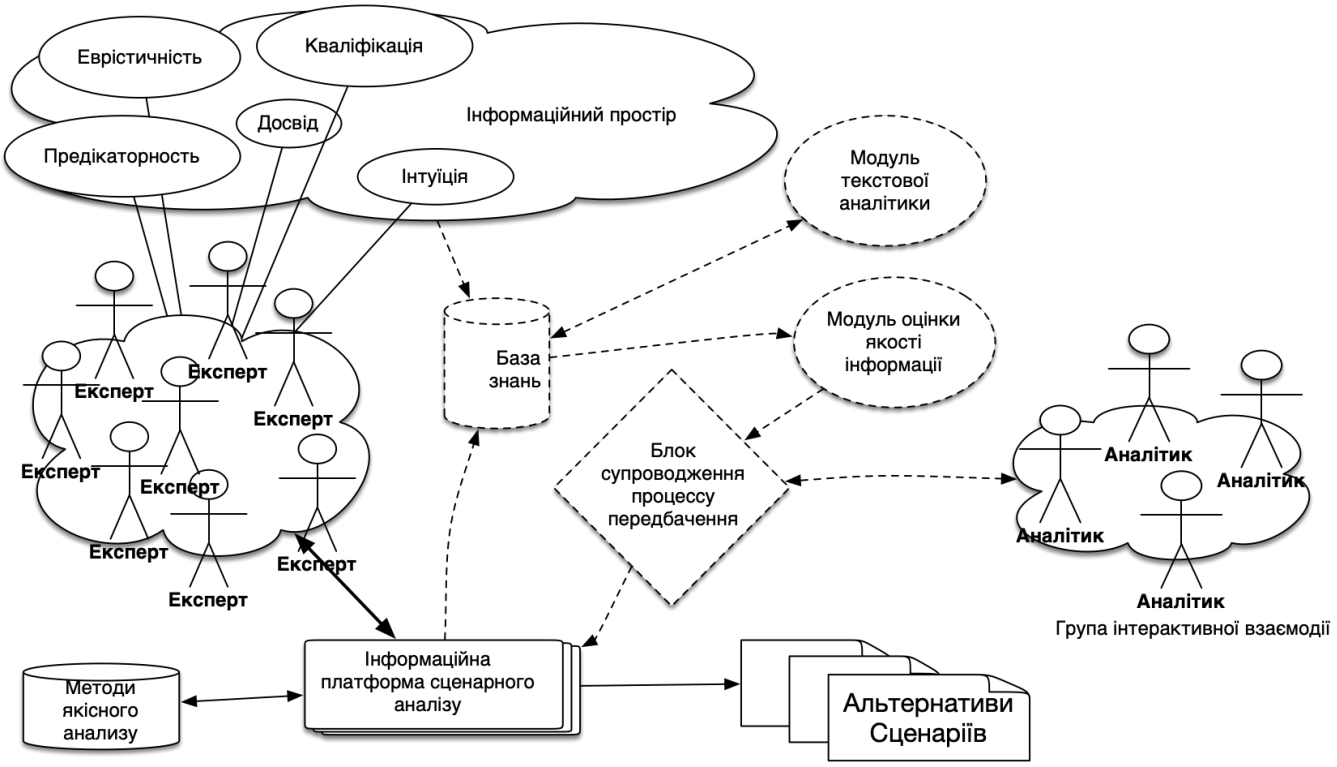

Рис. 3. Структурна схема модифікованої інформаційної моделі ПП

Введено класи метаданих в залежності від галузі чи предметного домену за призначенням, змістом, змістом із урахуванням емоційної забарвленості, вживанням з урахуванням часових параметрів та інтенсивності згадуваності.

Інформаційна модель предметної галузі включає статичний ієрархічній структурний компонент, а також організуючі зв'язки. Формалізований опис ієрархічної структури використовує теоретико-множинні поняття загальної теорії систем. У процесі аналізу досліджуваного об’єкта, суб’єкта або системи виникає його інформаційне відображення у вигляді конкретних, зафіксованих у даний проміжок часу метаданих. У формалізованому вигляді це відображення формує підмножину інформаційної моделі - класифікатор/категоризатор (один тип функціоналів) або онтологію предметної області (більш одного).

У разі використання класифікаторів та класифікуючих онтологій оптимізація щодо ергономічності представлення вже не є доцільною. У разі автоматизованого наповнення бази знань метаданими виникає задача оцінювання кількісної та якісної оцінки зібраних знань.

Перевагою використання деревовидних класифікаторів є простота відслідковування змін у структурі із плином часу та обсягів класифікованих знань і динаміки класифікування.

Згідно концептуальної моделі, у період часу t_0 існує набір показників інформованості $Q = <Q_1, Q_2, …., Q_i>$, що фіксують як стан структури категоризаторів, так і кількісно розмічені класифікатором знання і документи. У

кожний наступний момент часу t = t_1,...t_N як структура категоризаторів, так і структура знань, змінюються. З метою аналізу динаміки кількісних та якісних характеристик набуття знань введемо показники інформованості: відносно структури набутих знань, носіїв зібраної інформації, метаданих.

Для вилучення метаданих у процесі супроводження ПП було побудовано прийом вилучення фактів з текстів природною мовою на базі загальної моделі вилучення фактів з текстів природною мовою та апробовано і адаптовано до використання у інструментарії пакету текстової аналітики компанії SAS(R) та на базі бібліотек OpenSource. Модифікована модель із створеним набором правил для вилучення настроїв (сентиментів) наведена нижче:

$$E=<T , V , a> ,$$

де T - всі текстові об'єкти (документи) у вхідних даних, V - всі правила, a - логічна функція від (t_i, v_j), що приймає значення «Істина» якщо t_i відповідає v_j.

Відмінність пропонованої моделі від загальної моделі вилучення фактів з текстів природною мовою є в тому, що текст складає не послідовність слів, а послідовність абзаців, речень, а вже потім слів:

$$t = par_1 par_2 \ldots par_{|N|},$$
$$par = sent_1 sent_2 \ldots sent_{|M|},$$
$$sent = wrd_1 wrd_2 \ldots wrd_{|K|},$$
$$wrd_k \in \{Wrds, Pnkt, POS\_tag, * , Aa\},$$

де Wrds - це список слів, Pnkt - знаки пунктуації, POS_tag - теги частин мови, * - будь яке одне слово, Aa - будь-яке слово, що починається з великої літери.

Фрагмент тексту представляється схожим чином:

$$t=t_1 + t_2 + \ldots + t_j.$$

Необхідний набір фрагментів $T_{ri}^q = \{t\}$ для кожного тексту відповідає набору шаблонів правил r_i^q, і весь текст відповідає всім можливим шаблонам, фрагменти $T_r = U\, T_{ri}^q$ якого представлено наступим чином:

$$\forall p\ s_p \in s\, \forall t \in T^1_{ri\,s} \begin{cases} \exists w_{ij} \in \{w_i\},\ \forall j\, w_{ij} \in \{t\},\ j \geq 2, \\ |t| \geq 2, \\ \forall j\, w_{ij} \in c, \forall j\, w_{ij} \notin e,\ e = \oslash. \end{cases}$$

$$\forall p\ s_p \in s\, \forall t \in T^3_{ri\,s} \begin{cases} \exists w_{ij} \in \{w_i\},\ \exists j\ \exists sent_k\ w_{ij} \in sent_k,\ k \geq 1,\ j \geq 2, \\ t \in \{sent\}, |sent_k| \geq 2, \\ \forall j\, w_{ij} \in c, \forall j\, w_{ij} \notin e,\ e = \oslash. \end{cases}$$

$$\forall p\ s_p \in s\, \forall t \in T^4_{ri\,s} \begin{cases} \exists w_{ij} \in \{w_i\},\ \forall j\, w_{ij} \in \{t\},\ j \geq 2, \\ |t| \geq 2, \forall a, b \in \{j\}\ dist(w_{ia}, w_{ib}) \leq d, \\ \forall j\, w_{ij} \in c_1 \forall j\, w_{ij} \notin e,\ e = \oslash. \end{cases}$$

$$\forall p\ s_p \in s\, \forall t \in T^5_{ri\,s} \begin{cases} \exists w_{ij} \in \{w_i\},\ \forall j\, w_{ij} \in \{t\},\ j \geq 2, \\ |t| \geq 2, \forall a, b \in \{j\}, a < b, \\ \forall j\, w_{ij} \in c, \forall j\, w_{ij} \notin e,\ e = \oslash. \end{cases}$$

$$\forall p\ s_p \in s\, \forall t \in T^6_{ri\,s} \begin{cases} \exists w_{ij} \in \{w_i\},\ \forall j\, w_{ij} \in \{t\},\ j \geq 2, \\ |t| \geq 2, \forall a, b \in \{j\},\ dist(w_{ia}, w_{ib}) \leq d,\ a < b, \\ \forall j\, w_{ij} \in c, \forall j\, w_{ij} \notin e,\ e = \oslash. \end{cases}$$

$$\forall p\ s_p \in s\, \forall t \in T^7_{ri\,s} \begin{cases} \exists w_{ij} \in \{w_i\},\ \forall j\, w_{ij} \in \{t\},\ j \geq 3, \\ |t| \geq 3, \\ \forall a, b, c \in \{j\}, a < c < b, w_{ia}, w_{ib} \in c, w_{ic} \notin e. \end{cases}$$

$$\forall p\ s_p \in s\, \forall t \in T^8_{ri\,s} \begin{cases} \exists w_{ij} \in \{w_i\},\ \forall j\, w_{ij} \in \{t\},\ j \geq 1, \\ |t| \geq 1, \\ \forall j\, w_{ij} \notin e,\ e = \varnothing. \end{cases}$$

Шаблон r_i^q = <c, e, d>, де c - це лексичне обмеження, e - виняток з лексичного обмеження, d - задає межі покриття правилами, при цьому d ∈ N, q ∈ {1,2,3,4,5,6,7,8}.

Модифікація з урахуванням емоційно-семантичної орієнтації: остаточний набір правил вилучення фактів V= $\{v_i \mid i = \overline{1, N}\}$, має вигляд:

$$v_i = (\{<p_j, arg_j>\}, s, w),\ j \geq 1,$$

де ім'я аргументу $arg_j \in \{\varnothing, \{\div\ Z\}\}$, s - це сентімент (емоційно-семантична орієнтація), w - вага правила, s ∈ {-1; 0; 1}, w ∈ R (w ∈ (0; 10]). Крім того, можуть бути такі модифікації: s ∈ <∅, {+, =, -}, {-2; -1; 0; 1; 2}, {-3; -2; -1; 0; 1; 2; 3}> відповідно до здібностей розпізнавання людини.

$arg_j = \varnothing$ означає, що немає ніяких фактів, призначених для вилучення, достатньо тільки узгодження з шаблоном, тобто це - випадок класифікації (також і s = ∅). При $s \neq \varnothing$ можливо застосовувати правила для вилучення емоційної забарвленості.

Головна перевага пропонованої моделі над відомою є у принципі, за яким вилучається послідовність фактів. Згідно з визначенням, послідовність фактів може бути повернена у різні аргументи або послідовно витягнута та об'єднана в одному аргументі. При цьому префікси та постфікси можуть бути як наявні, так і відсутні навколо будь-якого факту, що вилучається.

Введені у модель лексичні обмеження-шаблони використовуються для створення в листах категоризаторів наборів більш складних правил.

За складністю побудови у системі обробки слабко структурованої інформації правила для класифікації можно умовно розділити на наступні (від найлегших до найскладніших): співставленням словникових термів без видобуття фактів; вилученням простих частково словникових фактів (показники, локація, ПІБ, час); вилученням складних фактів (співставлення фактів, порівняння); вилученням знань (семантичні конструкції цілепокладання, дефініції проблеми та ін., що перелічено в таблицях введених метаданих ПП); аналізом емоційної забарвленості. За допомогою правил з емоційною забарвленістю можна визначити позитивні, негативні або нейтральні тенденції у зовнішньому середовищі, наслідки можливих планів дій впливових суб'єктів, визначити проблеми. Для цього використовуються шість заздалегідь визначених концептуальних категорій для правил класифікації з можливістю вилучення фактів та визначення емоційної забарвленості, а саме: просте слово або фраза з емоційним забарвленням; зменшення або збільшення показника (властивості) досліджуваного об'єкту, суб'єкту або системи; високий, низький, зростаючий або спадаючий рівень потенційно негативного або позитивного показника; бажаний або небажаний факт; відхилення від норми або бажаного діапазону значень; генерація, споживання або втрата ресурсів.

У рамках засобів формування ключових понять правил та гілок класифікаторів досліджено та застосовано прийом для побудови первинної класифікуючої онтології із застосуванням латентного семантичного аналізу (LSA), що на сьогодні найчастіше за все використовує метод SVD (також, цей метод реалізовано у SAS(R) EM) та латентне розподілення Діріхле (bag-of-words).

Для генерації правил вилучення фактів (у тому числі, емоційно забарвлених) відносно об'єктів та їх властивостей використано прийоми, що базується на лексиконі та правилах: вилучення об'єктів та їх властивостей з використанням існуючого словаря позитивних або негативних слів; вилучення позитивних або негативних слів з використанням існуючої таксономії об'єктів та їх властивостей. На цьому етапі ідентифікація заперечення позитивних або негативних словосполучень не є суттєвим, тому що важливішим є вилучення самих значень об'єктів та їх властивостей, або позитивних чи негативних ознак. Вилучення об'єктів та їх властивостей з використанням існуючого словаря позитивних або негативних слів є автоматизованим процесом, проте сортування та подальша обробка виконується за допомогою системного аналітика.

Прийомом генерації правил з урахуванням емоційної забарвленості ситуацій та властивостей є визначення значимості іменних груп, що складають бажані та небажані факти. У даній роботі вперше введено ваговий коефіцієнт ω^s до формули загального балу $score$ (значущості іменних груп) при агрегації емоційного забарвлення:

$$score_k(f) = \sum_i \frac{SO_{w_i}}{dis(w_i, f)} \omega_k^s(e, d, L, D, t_j),\ i: < w_i,\ f > \in S \wedge w_i \in L_k,$$

$$\omega_k^s(e, d, L, D, t_j) = SF(e, d, L, D, t_j) * IDFS(L, D, t_j),$$

де w_i є емоційно-забарвленим емоцією $e^k \in e,\ e = < e^1, ..., e^k, ..., e^l, ..., e^K >$ словом, L_k - множина сентимент ознак (включаючи ідіоматичні висловлювання) кожної емоції з простору емоцій e^k, $k = \overline{1, K}$ - розмір обраного простору емоцій, S - номер речення, що містить властивість f, $dis(w_i,f)$ - дистанція між емоційно-забарвленим емоцією e^k словом w_i та властивістю f деякого досліджуваного об'єкту предметної галузі, SO_{w_i} - емоційно-семантична орієнтація сентимент-слова, $t_j \in [t_{start};t_{end}]$ - часовий інтервал, обмежуючий існування обраного простору емоцій, релевантного до набору впливових трендів у досліджуваній системі. Ваговий коефіцієнт ω^s_k розраховується як $SF*IDFS$ (аналогічно до метрики TF-IDF), проте новизною є те, що коефіцієнт розраховується не для слів (об'єктів предметної галузі), а для емоцій e^k з урахуванням плину часу:

$$SF^{e^l}(e, d, L, D, t_j) = \frac{n_{w_i \in L_l, d}}{\sum_1^K n_{w_i \in L_k, d}},\ w \in d,\ d \in D_{t_j},$$

$$IDFS^{e^l}(L, D, t_j) = \frac{|D_{t_j}|}{|\{d \in D_{t_j} : w_i \in L_l \wedge w_i \in d\}|}$$

де n_{w_i} - частота входження емоції e^l у документ, що визначено через кількість емоційно-забарвлених слів, визначаючих емоцію, $D_{t_j} \supset D$ підмножина корпусу D визначеного горизонту t_j.

Додатковою модифікацією розрахування коефіцієнту значимості став прийом визначення важливості потенційної властивості відносно інших, де ω^f_l розраховується аналогічно до ω^s_k:

$$score_k(f) = \sum_{w_i : w_i \in S \cap w_i \in L_k} \frac{w_i SO}{dis(w_i, f_l)} \omega^s_k(e, d, L, D, t_j) \omega^f_l(f, d, L, D, t_j).$$

Розглянуто прийом генерації правил для аналізу досліджуваної системи через високий, низький, зростаючий або спадаючий рівень потенційно негативного або позитивного показника. Він базується на виявленні станів “високий”, “низький”, “зростаючий” або “спадаючий” рівень потенційно негативного або позитивного показника властивості чи об’єкту. Для цього до множини показників додані типові показники з економічного лексикону:

$$kpi\_trends = kpi \times lvl \cap kpi \times dir,$$

де kpi = <arg_j>, $arg_j \in$ <*споживання, вартість, валюта, дефіцит, попит, знижка, надлишок, інвестиції, вихід, потужність, ціна, квота, ризик, частка, ринкова вартість, субсидія, поставка, тариф, податкова ставка, обсяг*>, lvl = <arg_k>, $arg_k \in$ <*великий, низький*>, dir = <arg_l>, $arg_l \in$ = <*ріст, спад*>.

Проте на часовому горизонті показники можуть трактуватися як потенційно позитивні чи негативні, в цілому же на часовій осі емоційно-семантична орієнтація їх може змінюватись. На практиці визначено 5 типів ситуацій неоднозначності через зміну емоційно-семантичної орієнтації та визначено прийоми їх розкриття автоматично у разі наявності видобутих фактів або за допомогою відповідного методу якісного аналізу. Наведені прийоми було апробовано на наступних предметних доменах: енергетична сфера; політика, економіка, суспільство, енергетика (інтерв'ю експерта); енергетика та конфлікт на сході.

Третій розділ присвячено моделям, алгоритмам та прийомам щодо реалізації системного підходу супроводження ПП з наявністю слабко структурованих даних засобами текстової аналітики у вигляді модулів інформаційної підсистеми платформи сценарного аналізу.

Приведено програмну реалізацію системи збору та збереження даних з джерел слабко структурованої інформації. На рис. 5 наведено структурну схему інформаційної моделі супроводження ПП.

На всьому життєвому циклі ПП модель отримує на вході слабко структуровані дані, категоризує їх, застосовує моделі вилучення знань та генерує на виході структуровані дані для методів якісного аналізу, висвітлює протиріччя у знаннях для автоматичного розкриття чи надає рекомендації щодо залучення методів якісного аналізу для усунення протиріч.

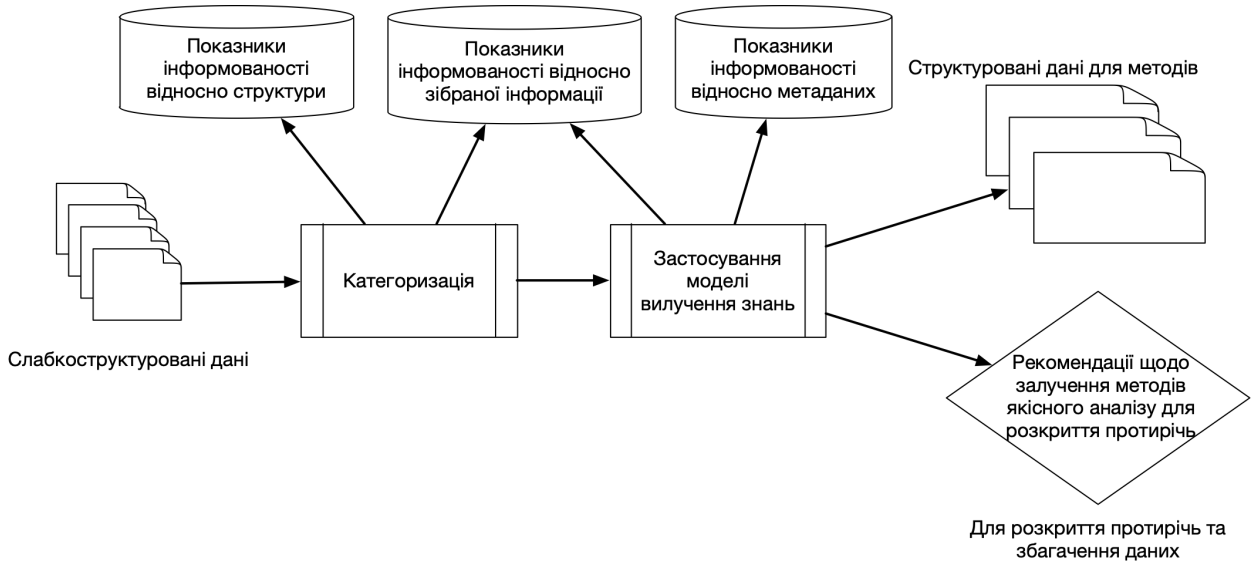

Рис. 5. Структурна схема інформаційної моделі супроводження ПП

У процесі оброблення інформації у моделі розраховуються та накопичуються набори даних у вигляді часових рядів із показниками інформованості відносно структури набутих знань, носіїв зібраної інформації та метаданих передбачення.

Розглянуто та класифіковано на чотири класи потенційні джерела неструктурованої інформації, серед яких на вході є: документи щодо аналізу стану досліджуваної системи; плани розвитку досліджуваної системи (у вигляді таблиці: проблема, захід, результат, бюджет); стенографовані аудіозаписи круглих столів та конференцій з питань розвитку; перелік інвестиційних проектів для створення та/або розвитку; профільні публікації та огляди стану системи або схожих систем; витяги із засідань Рад щодо системи або адміністративного чи законодавчого поля, що стосується об’єктів досліджуваної системи; плани розвитку об’єктів та підсистем схожих систем; план розвитку галузі у межах розглядуваної країни/регіону; паспорти передових інноваційних технологій; переліки класифікаторів, статистичні таблиці характеристик та показників об’єктів та підсистем розглядуваної системи; новини з медіа-ресурсів; блоги компаній-розробників у досліджуваній галузі; патенти; сторінки соціальних мереж; публікації твітеру; стенограми відеоматеріалів; інші джерела. Проведено аналіз щодо легальності зчитування змісту документів з джерел слабко структурованої інформації.

Задане первісне анотування елементів слабко структурованої інформації повинно здійснюватися до моменту надходження даних до ПП.

Наведено алгоритм процесу обробки вхідної інформації у рамках супроводження ПП, що реалізує запропоновану раніше та удосконалену структуру інформаційної платформи з додатковими блоками: бази знань (БЗ), модулем оцінки якості інформації (МЯІ) і супроводження процесу (БСП), текстової аналітики (ТА):

1. Отримання даних до інформаційної платформи передбачення.
2. Класифікація джерел. Видобуття первинних метаданих з якісних даних (data, author, source, та ін.).
3. Збір екземплярів інформації з кожного джерела.
4. Розміщення даних у базі знань.
5. Обробка вхідних даних у блоці текстової аналітики.

6. Запис видобутих даних, метаданих і знань до БЗ.
7. Передача інформації до блоку супроводження ПП.
8. Здійснення стратегії супроводження ПП при надходженні нових знань.
9. Надання інформації експертам і введення до бази знань додаткових зв'язків зі стенограм і інтерв'ю (brainstorm/mindmaps, cross impact, swot, morphological analysis, roadmap).
10. Перевірка повноти покриття знаннями на базі порівняння з покриттям класифікаторів та інших показників інформованості.
11. Перевірка повноти заповнення фреймової структури БЗ (indices, volume, speed, scale, power, amount, та ін).

Приведено узагальнену архітектуру платформи збору та збереження великих обсягів слабко структурованої інформації, що дозволяє гнучко масштабувати засоби обробки в залежності від інформаційної ємності вхідної інформації. До складу платформи входить: 1. Набір кравлерів; 2. Набір парсерів; 3. Платформа розподілення інформаційних потоків (RabbitMQ); 4. Платформа обробки/Сховище даних (Elasticsearch); 5. Підсистема візуалізації даних (Elasticsearch Kibana); 6. Аналітичний процес та NLP-процесор.

У **четвертому розділі** роботи приведена практична реалізація запропонованого системного підходу до супроводження ПП з наявністю слабко структурованих даних засобами текстової аналітики при розв’язанні проблем передбачення в різних проектах.

У рамках проектів “Інструментарій моделювання і сценарного аналізу планування розвитку інфраструктури мегаполісу в умовах екологічних, техногенних і терористичних загроз” та “Побудова інформаційно-аналітичної платформи сценарного аналізу на основі великих обсягів слабко структурованої інформації” було створено та застосовано ПЗ на базі OpenSource та пропрієтарного ПЗ у ході вирішення наступних підзадач: очищення корпусу; лематизація текстів корпусу (pymorphy2) з очищенням; побудова моделі Word2Vec (libgensim); вилучення концептуальних понять домену “Підземна та наземна інфраструктура мегаполісу”; вилучення концептуальних понять домену “COVID-19”; побудова класифікуючої онтології та генерація правил класифікації згідно її гілок; імплементація правил у SAS® Content Categorization Studio; завантаження моделі до SAS® Content Categorization Server; маркування (класифікація) текстів, вхідних та вихідних дані, способи представлення даних; адаптація підходу до великих об’ємів даних (на прикладі предметного домену COVID-19). Приведена реалізація дозволяє гнучко масштабувати засоби обробки в залежності від інформаційної ємності вхідної інформації.

В межах виконання проекту "Розроблення науково-методичного і програмного забезпечення виявлення перспективних напрямів розвитку новітніх технологій інноваційного розвитку на рівні великих підприємств, галузей та регіонів на основі технологічного передбачення" було використано широкий набір розроблених в дисертації методів, алгоритмів, прийомів згідно викладених моделей, а саме, вирішено наступні підзадачі: відбір та класифікація джерел; синтез правил класифікаторів та застосування існуючих класифікаторів; ідентифікація трендів галузі енергоринку через видобуття фактів про високий / низький або що росте /

спадає рівнях потенційно позитивного або негативного показника; порівняння станів та трендів у галузі енергоринку у динаміці часу; аналіз конфліктів знань через динаміку та стан рівня потенційно позитивного чи негативного показника; ідентифікація ключових об’єктів/актуальних проблем галузі Енергетика через вилучення емоційного забарвлення із зважуванням емоційного фону (через коефіцієнт значимості емоції); виявлення ключових технологій через аналіз інтерв’ю/звіту експерта; обчислення показників інформованості бази знань передбачення. В результаті виконання цього проекту був оброблений масив текстових документів, з якого були вилучені в БЗ: 25360 об'єктів, 406 об'єктів предметної області енергетика, 11191 об'єктів-учасників трендів, 2000 об'єктів-учасників проблем, 1862 об'єкта в цілях, 378 технологій, 225 проблем, 1385 трендів, 112 цілей; було виділено 12 важливих трендів, 143 проблеми, 82 мети, 253 технології, що було використано у методах якісного аналізу.

У проекті «Розробка інформаційно-аналітичних засобів дослідницької служби у складі інтегрованої інформаційно-аналітичної системи “Електронний Парламент”» було застосовано модель та підхід вилучення знань з текстів природною мовою для визначення перехресного впливу урядових заходів на види економічної діяльності.

При виконанні проекту Modeling and mitigation of social disasters caused by catastrophes and terrorism (NATO SPS G4877)” було перевірено на практиці прийом щодо генерації правил класифікатора надзвичайних явищ ДК 019:2010 як вже існуючого класифікатору державного значення. Для дослідження явища корупції та висвітлення у ЗМІ трендів по боротьбі із корупцією як фактору впливу на пом'якшення соціальних лих було сформовано класифікуючу онтологію з 4 класів та правила класифікації у вигляді лексичних обмежень згідно моделі вилучення знань з текстів природною мовою, проведено класифікацію текстів та аналіз результатів.

ВИСНОВКИ

У дисертаційній роботі вирішена проблема розробки математичного забезпечення супроводження процесу передбачення (СПП) з наявністю слабко структурованих даних засобами текстової аналітики, що застосовується при вирішенні практичних задач на рівні підприємств, відомств, галузей, регіонів. Основні наукові та практичні результати роботи полягають в наступному:

Проаналізовано існуючі підходи до СПП, визначено базові інформаційні одиниці - метадані. Визначено недоліки існуючої моделі СПП.

Запропоновано концепцію СПП на базі конусу часу. Розглянуто ступінь невизначеності та швидкість плину часу відносно складної системи з людським фактором. Сформовано фактори, що розширюють конус за виміром невизначеності у часі T(N).

Розроблено системний підхід до СПП засобами текстової аналітики на основі прийомів та алгоритмів обробки слабко структурованих даних.

Запропоновано модифіковану інформаційну модель ПП та введено додаткові метадані. Створено інформаційну модель предметної галузі. Наведено ієрархічне представлення досліджуваної системи як класифікуючої онтології. Розглянуто проблематику представлення знань у вигляді онтології та визначено доцільність

використання класифікуючих онтологій - що реалізують ієрархічну деревоподібну структуру з одним відношенням-функціоналом.

Створено концептуальну модель якості знань та введено інтегровані показники інформованості в залежності від часу у трьох вимірах відносно: структури набутих знань, носіїв зібраної інформації, метаданих модифікованої інформаційної моделі ПП.

Розроблено модель вилучення фактів та знань із слабко структурованих даних на базі існуючої загальної моделі вилучення фактів з текстів природною мовою. Модель базується на більш детальному представленню фрагментів тексту та на створених 8 шаблонах, що є лексичними обмеженнями та основою правил-фільтрів для вилучення знань з предметної області у вигляді метаданих модифікованої інформаційної моделі передбачення.

Розроблено модель для врахування емоційної забарвленості через вилучення позитивних чи негативних ознак.

Запропоновано наступні прийоми щодо вилучення об’єктів-метаданих інформаційної моделі передбачення та їх властивостей.

Досліджено ситуацію виникнення конфліктів знань та створено прийоми щодо їх розв'язання. Введено ваговий коефіцієнт значущості іменних груп, що складають бажані та небажані факти, у тому числі з урахуванням часу життя об’єктів у інформаційному потоці на вході передбачення. Окреслено ситуації зміни емоційно-семантичної орієнтації та наведено неоднозначності та конфлікти знань, що виникають як наслідок таких ситуацій. Розглянуто підходи щодо автоматизованого та експертного усунення ситуацій неоднозначності та конфлікту знань.

Проведено апробацію системного підходу до супроводження процесу передбачення з наявністю слабко структурованих даних засобами текстової аналітики.

Створено обчислювальні модулі реалізації прийомів обробки слабко структурованих даних і текстової аналітики. Приведено приклад програмної реалізації системи збору та збереження даних з джерел слабко структурованої інформації, що дозволяє гнучко масштабувати засоби обробки в залежності від інформаційної ємності вхідної інформації. Застосування і ефективність запропонованих підходів, моделей та алгоритмів ілюструються можливостями щодо автоматизованої обробки великих обсягів слабко структурованих даних в проектах дослідження розвитку систем із людським фактором.

Приведено застосування системного підходу на прикладі задач проектів «Розробка інформаційно-експертної системи передбачення з урахуванням поглибленої аналітики неструктурованих даних», «Розробка інформаційно-аналітичних засобів дослідницької служби у складі інтегрованої інформаційно-аналітичної системи “Електронний Парламент”» та у проекту Modeling and mitigation of social disasters caused by catastrophes and terrorism (NATO SPS G4877). В рамках поставлених задач проілюстровано роботу методів, моделей та алгоритмів видобуття фактів та приклади супроводження методів якісного аналізу необхідними знаннями.

Розроблений системний підхід застосовується на всьому життєвому циклі сесії передбачення та забезпечує зменшення ресурсів забезпечення у внутрішніх підпроцесах системи та покращує якість процесів, а саме: прискорює обробку вхідних даних ПП, забезпечує аналітиків та експертів засобами швидкого аналізу вхідних даних у ході ПП, інформацією про хід ПП у вигляді показників інформованості, забезпечує повторне використання видобутих знань та здобутих артефактів на виході моделей, алгоритмів та підходів у наступних сесіях передбачення. Розв'язання низки практичних задач підтвердило результативність, ефективність, масштабність запропонованої концепції цілісності ПП при залученні запропонованого системного підходу.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

У монографіях:

1. Панкратова Н. Д. Моделирование альтернатив сценариев процесса технологического предвидения / Н. Д. Панкратова, В. В. Савастьянов // Инновационное развитие социо-экономических систем на основе методологий предвидения и когнитивного моделирования / Под ред. Гореловой Г.В., Панкратовой Н.Д. – Киев: Наукова думка. -2015. – С 344-360, опублікованій у співавторстві, здобувачеві належать: інформаційна модель передбачення, підходи щодо моделювання та супроводження альтернатив сценаріїв.

У наукових фахових виданнях:

1. Pankratova N.D. Foresight Process Based on Text Analytics / Pankratova N.D., Savastiyanov V.V. // International Journal «Information Content and Processing». — 2014. — 1, No 1, ITHEA. — P. 54–65., входить до наукометричних баз Worldcat, ROAD, Google Scholar, CiteseerX, ITHEA, опублікована у періодичних наукових виданнях інших держав, які входять до Європейського Союзу, у співавторстві, здобувачеві належать: нові метадані процесу передбачення, інформаційна модель процесу передбачення, алгоритм процесу обробки вхідної інформації, модифікована модель вилучення фактів з текстів.
2. Pankratova N. D. Foresight and Forecast for Prevention, Mitigation and Recovering after Social, Technical and Environmental Disasters / N. D. Pankratova, P. I. Bidyuk, Y. M. Selin, I. O. Savchenko, L. Y. Malafeeva, M. P. Makukha, V. V. Savastiyanov // Springer. — 2014. — P. 119-134., входить до наукометричних баз SCOPUS, Web of Science, Google Scholar, опублікована у періодичних наукових виданнях інших держав, які входять до Європейського Союзу, опублікованій у співавторстві, здобувачеві належать: метод обробки слабко структурованих даних у формуванні альтернатив сценаріїв.
3. Savastiyanov V.V. Development of tools for analysis of texts of public and specialized sources in the tasks of prediction and system analysis. System Research&Information Technologies, №4, входить до наукометричних баз SCOPUS, DOAJ, Index Copernicus, РИНЦ та ін, входить до фахових видань України категорії “Б” - 2020.- P.10-23

4. Панкратова Н. Д. Моделирование альтернатив сценариев процесса технологического предвидения / Н. Д. Панкратова, В. В. Савастьянов // Системні дослідження та інформаційні технології, входить до наукометричних баз DOAJ, Index Copernicus, РИНЦ та ін, входить до фахових видань України категорії “Б” — 2009. — № 1. — С.22–35, опублікованій у співавторстві, здобувачеві належать: інформаційна модель, зручною для подання в пам'яті ЕОМ, що утворює базу і поле знань, побудована на основі мережі фреймів; стратегія інформаційного моделювання альтернатив сценаріїв.
5. Савастьянов В. В. Технологическое предвидение информационно-компьютерных технологий связи / В. В. Савастьянов // Системні дослідження та інформаційні технології, входить до наукометричних баз DOAJ, Index Copernicus, РИНЦ та ін, входить до фахових видань України категорії “Б”— 2005.
6. Терентьєв О. М. Застосування когнітивного та ймовірнісного моделювання в задачах формування сценаріїв розвитку соціально-економічних систем / О. М. Терентьєв, Т. І. Просянкіна-Жарова, В. В. Савастьянов // Наукові вісті НТУУ “КПІ”. – №5, входить до наукометричних баз DOAJ, Index Copernicus, РИНЦ та ін, входить до фахових видань України категорії “Б”, – К.: НТУУ “КПІ” ВПІ ВПК “Політехніка”, 2016. – 37-47 с. – DOI: http://dx.doi.org/10.20535/1810-0546.2016.5.79876, у співавторстві, здобувачеві належать: синтез правил обробки вхідних даних у слабко формалізованому вигляді для вилучення факторів, концептів, причинно-наслідкових зв’язків.

У інших виданнях:

1. Терентьєв О.М. Використання засобів текстової аналітики як інструменту оптимізації підтримки прийняття рішень у задачах розробки планів соціально-економічного розвитку України / О.М. Терентьєв, Т. І. Просянкіна-Жарова, В. В. Савастьянов // Реєстрація зберігання та обробка даних. – Т. 18. – № 3. – К.: ТОВ “Інфодрук”, 2016. – 75-86 с. – ISSN 1560-9189, у співавторстві, здобувачеві належать: інформаційно-лексична модель соціально-економічної системи для категоризації даних, підходи текстової аналітики для вилучення трендів, фактів росту/падіння потенціально позитивного/негативного показника.
2. Савастьянов В. В. Построение информационной модели сопровождения процесса технологического предвидения / В. В. Савастьянов // Наукові праці. Комп'ютерні технології : науково-методичний журнал. - Миколаїв: Видавництво МДГУ ім. Петра Могили, 2008, т.90 N 77, С.80-86.
3. Савастьянов В.В. Стратегія технологічного передбачення при моделюванні ринків телекомунікації / В. В. Савастьянов // Наукові праці: Науково-методичний журнал. — Т. 68. Вип. 55. Комп'ютерні технології. — Миколаїв: Вид-во ЧДУ ім. Петра Могили, 2004. — С.62–68.

Список патентів здобувача:

1. Згуровський М. З. Патент UA № 22435, МПК (2006) G06Q 10/00, ІНФОРМАЦІЙНО-АНАЛІТИЧНА СИСТЕМА ЗБОРУ ТА ОБРОБКИ ДАНИХ

/ М. З. Згуровський, Н. Д. Панкратова, А. М. Радюк, П. В. Будаєв, В. В. Савастьянов, Е. С. Клименко // Заяв. 13.11.2006, Опубл. 25.04.2007, бюл. № 5/2007, у співавторстві, здобувачеві належать: алгоритм імпорту даних, що реалізує пошук пов'язаної інформації із зовнішніх джерел в режимі автоматичного пошуку за критеріями автоматичної агрегації з відомих джерел та напівавтоматичної агрегації з інших джерел інформації.

Дисертантом робились доповіді на міжнародних наукових конференціях:

1. Савастьянов В.В. Моделирование ранних этапов процесса технологического предвидения. / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали IX Міжнародної науково-технічної конференції. — К.: ННК «ІПСА» НТУУ «КПІ», 2007.
2. Савастьянов В.В. Информационная модель сопровождения процесса технологического предвидения. / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали X Міжнародної науково-технічної конференції. — К.: ННК «ІПСА» НТУУ «КПІ», 2008.
3. Савастьянов В.В. Построение информационной модели задач технологического предвидения. / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали XI Міжнародної науково-технічної конференції. — К.: ННК «ІПСА» НТУУ «КПІ», 2009.
4. Савастьянов В.В. Моделирование процесса технологического предвидения. / Савастьянов В.В. // Информационно-компьютерные технологии в экономике, образовании и социальной сфере: тезисы докладов V всеукраинской научно-практической конференции. — Симферополь: КРП "Видавництво "Кримнавчпеддержвидав"", 2010, ISBN 978-966-354-352-9
5. Савастьянов В.В. Моделирование процесса технологического предвидения. / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали 12-ї Міжнародної науково-технічної конференції SAIT-2010. — К.: ННК «ІПСА» НТУУ «КПІ», 2010. ISBN 978-966-2153-41-5.
6. Савастьянов В.В. Моделирование и информационное сопровождение процесса предвидения / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали 13-ї Міжнародної науково-технічної конференції SAIT-2011. — К.: ННК «ІПСА» НТУУ «КПІ», 2011. ISBN 978-966-2153-41-5.
7. Савастьянов В.В. Ассоциативный анализ предпочтений посетителей веб-ресурсов в SAS® Enterprise MinerTM / Савастьянов В.В., Макуха М.П., // Системний аналіз та інформаційні технології: Матеріали 14-ї Міжнародної науково-технічної конференції SAIT-2011. — К.: ННК «ІПСА» НТУУ «КПІ», 2011. ISBN 978-966-2153-41-5.
8. Савастьянов В.В. Подход к информационному сопровождению процесса предвидения / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали 15-ї Міжнародної науково-технічної конференції SAIT-2014. — К.: ННК «ІПСА» НТУУ «КПІ», 2012. ISBN 978-966-2153-41-5
9. Савастьянов В.В. Стратегия моделирования процесса сценарного анализа / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали

13-ї Міжнародної науково-технічної конференції SAIT-2011. — К.: ННК «ІПСА» НТУУ «КПІ», 2011. ISBN 978-966-2153-41-5.

10. Savastiyanov V.V. Discovering of potential positive and negative factors of social disaster using sentiment analysis / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали 17-ї Міжнародної науково-технічної конференції SAIT-2015. — К.: ННК «ІПСА» НТУУ «КПІ», 2015. ISBN 978-966-2153-41-5
11. Терентьєв О.М. Текстовая аналитика в антикоррупционной деятельности / Терентьев А. Н., Савастьянов В. В., Макуха В. П., Просянкина-Жарова Т. И.// Научная конференция "Интеллектуальный системы в информационном противоборстве", 8-11 декабря 2015 г., Москва. – М.: ФГБОУ ВО "РЭУ" им. Г.В. Плеханова, 2015. – С. 220-224. – ISBN 978-5-7307-1064-1.
12. Terentiev O.M Analysis and modeling the dynamics changing of registered crimes taking into account the macroeconomic and political situation in Ukraine / Terentiev O.M., Makukha M.P., Savastynov V.V., Oparina E.L. // Системний аналіз та інформаційні технології: матеріали 18-ї Міжнародної науково-технічної конференції SAIT 2016, Київ, 30 травня – 2 червня 2016 р.– К.: ННК "ІПСА" НТУУ "КПІ", 2016. – С. 318-319.
13. Бідюк П.І. Застосування інструментів SAS Base для дослідження ефективності методів обробки пропусків у вибірках даних з метою підвищення якості прогнозування показників продовольчої безпеки країни / Бідюк П.І., Терентьєв О.М., Просянкіна-Жарова Т.І., Савастьянов В. В. // Системний аналіз та інформаційні технології: матеріали 19-ї Міжнародної науково-технічної конференції SAIT 2017, Київ, 22-25 травня 2017 р.– К.: ННК "ІПСА" НТУУ "КПІ", 2017. – С. 253-254. – ISBN 978-966-2748-94-9
14. Pankratova N., Savastiyanov V. Assessment of situations in the field of social disasters basing on the methodology of foresight and textual analytics. Proceedings of the 2019 IEEE Second International Conference IEEE UKRCON-2019 p. 1207-1210, ISBN 9781728138831

АНОТАЦІЯ

Савастьянов В. В. Супроводження процесу передбачення з наявністю слабко структурованих даних засобами текстової аналітики. - На правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 01.05.04 «Системний аналіз і теорія оптимальних рішень» (124–Системний аналіз). – Інститут прикладного системного аналізу Національного технічного університету України "Київський політехнічний інститут імені Ігоря Сікорського", Київ, 2021.

У роботі пропонується розглядати процес передбачення з наявністю слабо структурованих даних цілісно. Розроблено системний підхід до супроводу процесу передбачення на основі засобів текстової аналітики з чотирьох етапів, які безперервно повторюються на всьому життєвому циклі передбачення, а його результати використовуються повторно в рамках всіх інших сесій супроводу

процесів передбачення. На першому етапі визначаються моделі, методи і їх метадані, які будуть використовуватися для подання предметної області. Вводиться інформаційна модель процесу передбачення, інтегровані показники інформованості. Безперервно розраховуються і аналізуються показники інформованості. На другому етапі вводиться і застосовується модель і прийоми вилучення знань з текстів природною мовою. Розглянуто ситуації конфліктів знань і прийоми до їх усунення. На третьому етапі вводиться інформаційна модель супроводу процесу передбачення, вхідні / вихідні дані, алгоритм, який реалізує модель. На четвертому етапі проводиться адаптація і масштабування системного підходу.

Використання зазначеного системного підходу забезпечує зменшення ресурсів, необхідну для забезпечення даними внутрішніх подпроцессов, і покращує якість процесів, а саме: прискорює обробку вхідних даних процесу передбачення, забезпечує аналітиків і експертів засобами швидкого аналізу вхідних даних, інформацією в вигляді показників інформованості, забезпечує повторне використання здобутих знань та отриманих артефактів на виході моделей, алгоритмів і підходів в наступних сесіях передбачення.

Ключові слова: системний аналіз, методологія передбачення, текстова аналітика, natural language processing, data mining, супроводження процесу передбачення, сентимент аналіз, показники інформованості передбачення, інформаційна модель, концептуальна модель, модель вилучення знань з текстів природною мовою, класифікатори, синтез правил класифікації, метадані процесу передбачення.

АННОТАЦИЯ

Савастьянов В. В. Сопровождение процесса предвидения с наличием слабо структурированных данных средствами текстовой аналитики. - На правах рукописи.

Диссертация на соискание ученой степени кандидата технических наук по специальности 01.05.04 «Системный анализ и теория оптимальных решений» (124-Системный анализ). - Институт прикладного системного анализа Национального технического университета Украины "Киевский политехнический институт имени Игоря Сикорского", Киев, 2021.

В работе предлагается рассматривать процесс предвидения с наличием слабо структурированных данных целостно. Разработан системный подход к сопровождению процесса предвидения на основе средств текстовой аналитики из четырех этапов, которые непрерывно повторяются на всем жизненном цикле предвидения, а его результаты используются повторно в рамках всех других сессий сопровождения процессов предвидения. На первом этапе определяются модели, методы и их метаданные, которые будут использоваться для представления предметной области. Вводится информационная модель процесса предвидения, интегрированные показатели информированности. Непрерывно рассчитываются и анализируются показатели информированности. На втором этапе вводится и применяется модель и приемы извлечения знаний из текстов на естественном языке. Рассмотрены ситуации конфликтов знаний и приемы к их устранению. На третьем

этапе вводится информационная модель сопровождения процесса предвидения, входные/выходные данные, алгоритм, реализующий модель. На четвертом этапе проводится адаптация и масштабирование системного подхода.

Использование указанного системного подхода обеспечивает уменьшение ресурсов, требуемое для обеспечения данными внутренних подпроцессов, и улучшает качество процессов, а именно: ускоряет обработку входных данных процесса предвидения, обеспечивает аналитиков и экспертов средствами быстрого анализа входных данных, информацией в виде показателей информированности, обеспечивает повторное использование добытых знаний и полученных артефактов на выходе моделей, алгоритмов и подходов в следующих сессиях предвидения.

Ключевые слова: системный анализ, методология предвидения, текстовая аналитика, natural language processing, data mining, сопровождение процесса предвидения, сентимент анализ, показатели информированности предвидения, информационная модель, концептуальная модель, модель извлечения знаний из текстов на естественном языке, классификаторы, синтез правил классификации, метаданные процесса предвидения.

ANNOTATION

Savastiyanov V. V. Supporting foresight using textual analytics for semistructural data. - Manuscript copyright.

The thesis for candidate degree of technical science on speciality 01.05.04 – "System analysis and the theory of optimal solutions" (124–System analysis). – National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" MSE of Ukraine, Kyiv, 2021.

The paper proposes to review the process of foresight with the presence of semistructured data as a whole, gradually reducing uncertainty, moving from the start of the study to the desired future. To implement the proposed concept, a systematic approach to the support of the foresight process based on textual analytics, which is the most modern and most powerful tool for the analysis of semistructured data written in natural language.

The system approach consists of four stages which are continuously repeated throughout the life cycle of foresight, and its results are reused in all other foresight sessions. In the first stage, the subject area is studied, the features to the desired future are analysed, the models, methods and their metadata are determined. The conceptual model of support of the foresight process is determined. An idea of the process of foresight and the horizon of foresight is formed. Factors of growth and reduction of uncertainty on the way to the forecast horizon are determined. An information model of the foresight process is introduced - the representation of subject areas using the set-theoretic concept of general systems theory. Restrictions on information model connections are introduced, options for presenting knowledge in the form of a hierarchical classifier or ontology are considered, and advantages and disadvantages are outlined. The concept of the existence of knowledge in time is considered. Integrated time-dependent awareness indicators have been introduced to measure changes in the knowledge base over time and / or depending on the amount of new knowledge. New knowledge is registered as classified metadata according

to developed classifiers. Awareness indicators are constantly calculated and analyzed during the foresight process.

At the second stage of the system approach the model and approach of extraction of knowledge from texts in natural language is introduced and applied. The work modifies the general model of extracting facts from texts in natural language to meet the requirements of extracting metadata information model of foresight, introducing universal lexical templates-restrictions to compile more powerful rules for extracting metadata. The model is used as part of the support process to build techniques, approaches and tools based on them to process new subject areas and types of knowledge. Approaches to the extraction of objects of the subject area for the construction and expansion of classifiers, as well as approaches to generate classification rules for classifier nodes. Introduced methods for processing facts that contain potentially positive and negative indicators, including taking into account the time and changes in context. Situations of knowledge conflicts due to changes in emotional and semantic orientation and approaches to their elimination are considered.

At the third stage of the system approach the information model of support of the foresight process is introduced, classes of input data are defined. Metadata for the initial annotation and metadata to support the foresight process are introduced. The algorithm of transformation of input data into metadata, indicators of awareness and cases of usage of certain methods of qualitative analysis for sake of eliminating contradictions of knowledge in the knowledge base are presented. The data at the output of the foresight support process and the possibilities for their application at different stages of foresight and in the methods of qualitative analysis are considered.

At the fourth stage of the system approach, the semistructured data processing modules are adapted and scaled as a part of the foresight process support system. A number of cases show the application of a systematic approach to support the foresight process with the presence of semistructured data using textual analytics.

The developed system approach is applied throughout the life cycle of the foresight session. Artifacts created at the end of the support process (classifiers, lexical restrictions, rules, knowledge) can be used in subsequent and new foresight sessions.

Introduced system approach reduces the resources to provide data in the internal subprocesses of the system and improves the quality of processes, including: speeds up the processing of input data about foresight process, provides analysts and experts with tools for rapid analysis of input data, information on the progress in the form of awareness indicators, provides reuse of acquired knowledge and artifacts at the output of models, algorithms and approaches in subsequent foresight sessions. Number of practical cases confirmed the effectiveness, efficiency, scale of the proposed concept, saving the integrity of the foresight process, during the involvement of the proposed system approach.

Keywords: systems analysis, foresight methodology, text analytics, natural language processing, data mining, foresight process support, sentiment analysis, foresight awareness indicators, information model, conceptual model, model of knowledge extraction from texts in natural language, classifiers, synthesis of classification rules, foresight process metadata.